

## *Predictive Analytics – 1*

### HR Analytics Case Study

#### Group Members:

1. Ranjidha Rajan - (DDA1710198)
2. Soumadiptya - (DDA1710089)
3. Nihar Behera - (DDA1710367)
4. Anugraha Sinha - (DDA1710381)

# The Problem Statement

## ❑ Some statistics on XYZ company

- Current Workforce strength ~ 4000 employees
- Each year nearly 15% of employees leave the company
- This attrition rate is much higher than the ideal turnover rate of ~10%.

### 2013 Total Employee Turnover Rate by Industry (U.S.)

All Industries	15.1%
Banking & Finance	17.2%
Healthcare	16.8%
Hospitality	29.3%
Insurance	10.4%
Manufacturing & Distribution	13.3%
Not-for-Profit	15.3%
Services	15.2%
Utilities	7.2%

Source: <http://www.compensationforce.com/2014/02/2013-turnover-rates-by-industry.html>

However, it turns out that 10% is the **golden turnover number** you should try to aim for.

According to a 2013 survey, industry vertical wise attrition rate in the US

Delay in Projects making it difficult to meet timelines resulting in a reputation loss among customers and partners

Maintenance cost behind large department for recruiting new talent

Problems created by a High Attrition rate

Training new employees incurs additional costs

Time involved in acclimatizing for new employees and getting used to company culture leads to further delays in meeting timelines.

# Goals and Methodology



Management of XYZ company have contracted our firm to figure out

- Most important factors responsible for Attrition
- Provide suggestions to reduce the Attrition rate
- Pinpoint and make changes to their workplace to get most employees to stay



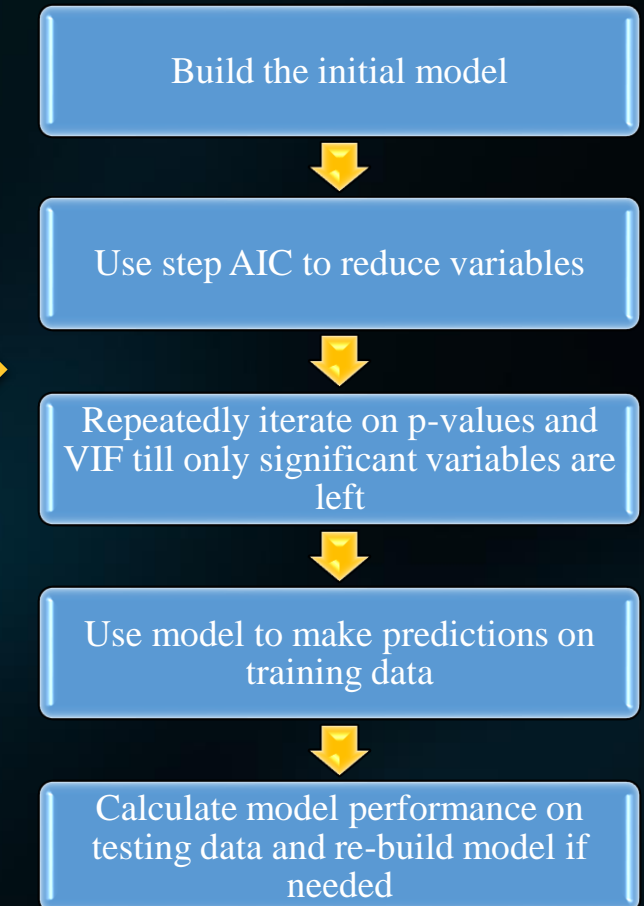
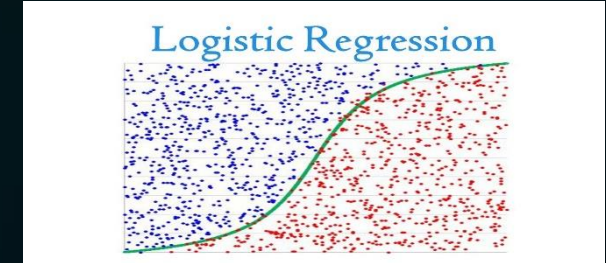
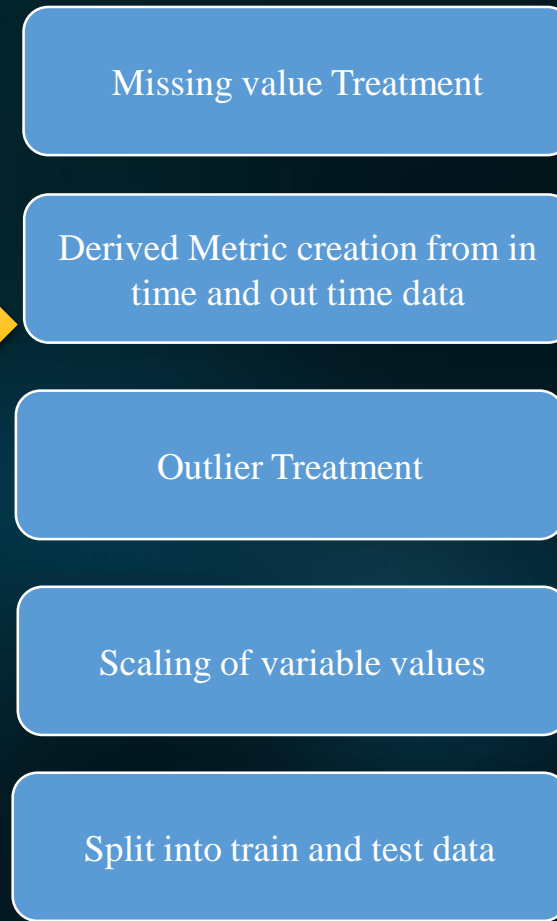
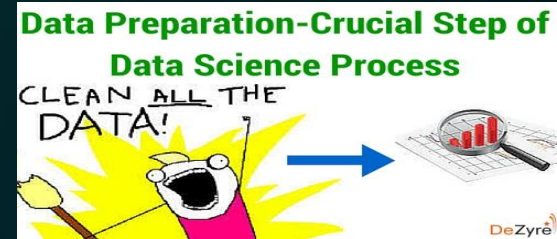
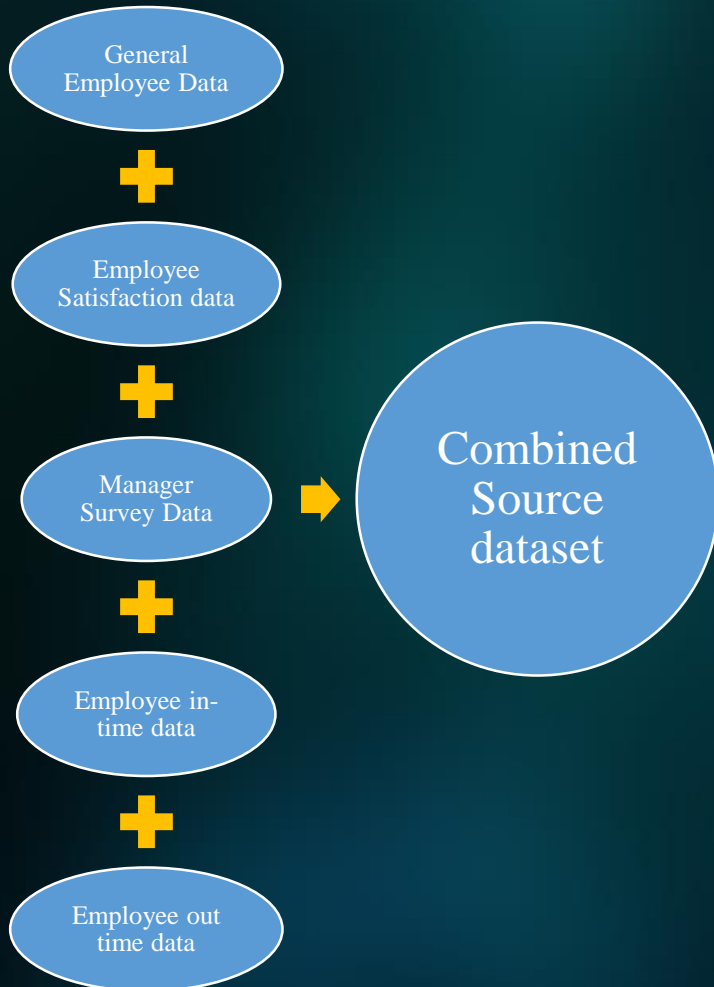
Our Problem solving methodology:-

- Model Probability of Attrition using a Logistic Regression model
- Figure out the most important variables from the Model
- Use the variables and their co-efficients to infer how they are related to the Attrition rate.
- Suggest how to curb the Attrition rate based on the findings.

# Data Analysis Methodology

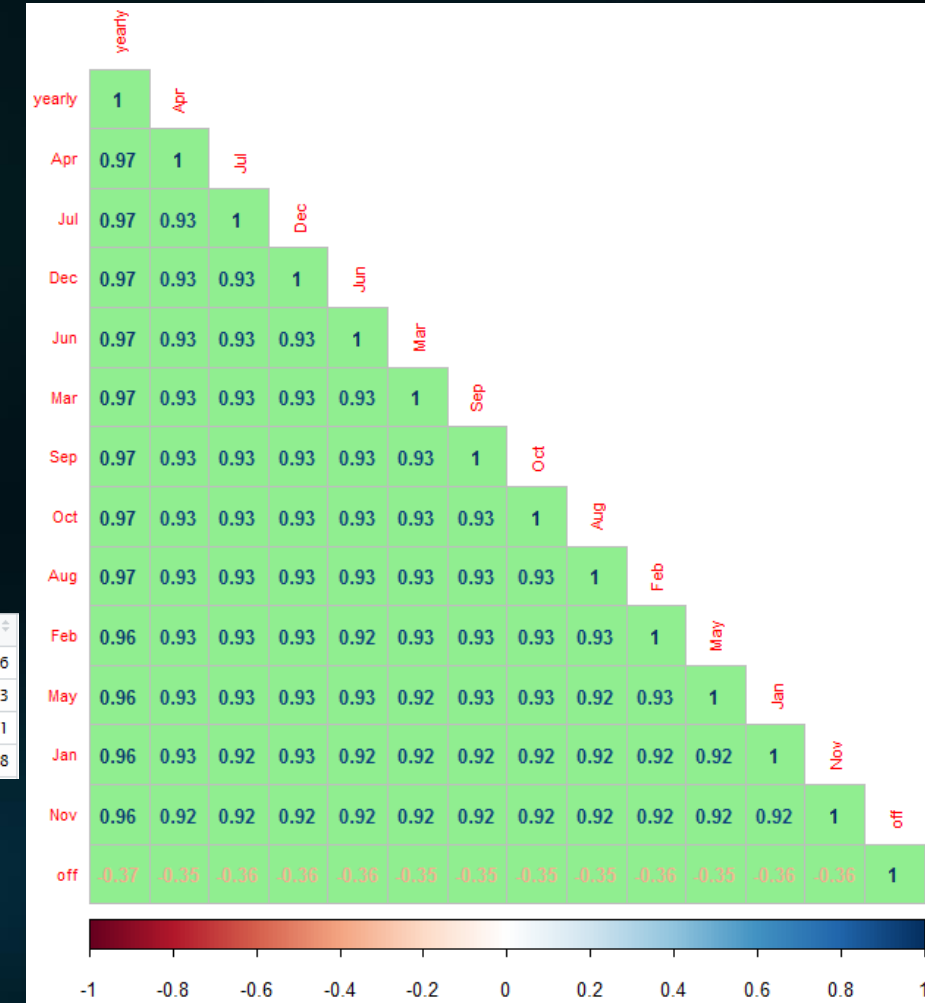


## DATA SOURCING



- ❑ Working on in\_time and out\_time data.
  - ❑ Total columns : 250 (249 days + 1 Employee ID)
  - ❑ Total rows : 4410
  - ❑ Number of columns with all NAs (holidays) : 12 days
  - ❑ Data processing :
    - ❑ For every employee build a monthly (percentage) working hours.
    - ❑ For every employee build a total number of yearly offs
    - ❑ For every employee build a yearly (percentage) working hours

EmployeeID	1	2	3	4	5	6	7	8	9	10	11	12	totalOffs	percYearlyWorking
1	0.8607602	0.8811111	0.8307970	0.9297270	0.7866528	0.7944665	0.9177399	0.9249686	0.8791121	0.7900033	0.9132022	0.8019760	17	0.8587786
2	0.8879989	0.9246094	0.9113145	0.9217519	0.8687448	0.9252794	0.8785243	0.9709921	0.9159540	0.9173264	0.9169946	0.9306850	13	0.9144963
3	0.8776224	0.8659271	0.7940972	0.8497932	0.8663177	0.8350758	0.8864347	0.8222338	0.8872867	0.8726042	0.8282677	0.8387563	7	0.8520101
4	0.8823337	0.8460017	0.7696445	0.8515325	0.8632413	0.9061995	0.8720312	0.7739087	0.8565030	0.8071776	0.8490471	0.9034343	14	0.8486518

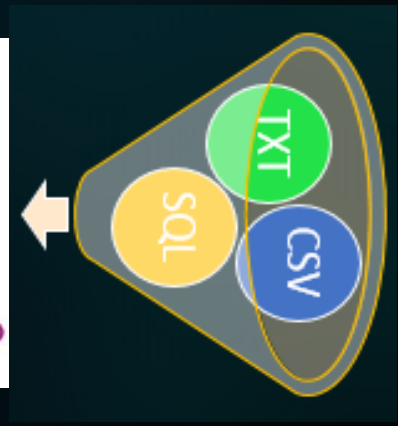


Note:  
Rather than taking mean as a metric for measuring this pattern, percentage has been chosen as a more viable metric, as comparison by percentage is more absolute for pattern reflection in this case.

- ❑ Build a correlation between different columns to see if monthly data for each employee helps in building some patterns
- ❑ Conclusion:
  - ❑ Monthly data for employees are highly correlated
  - ❑ Yearly percentage working hours is a good figure for reflecting employee's working hours.
  - ❑ Yearly Offs is also a good metric for reflecting employee patterns.

## □ Data Understanding

- Employee survey (3 features) : Survey of employee satisfaction on different parameters
- Manager survey (2 features) : Employee opinion of managers
- In time and out time data (Daily) : Employee's in-time and out-time data (Period : 12 months)
- General Data (24 features) : Various information about employees, department, Job Role etc.
- Primary key linking all data : Employee ID (Total employee list – 4410 employees)



Strategy	Remarks	Execution and Results	Number of columns found
Single Value Type Columns	Columns where count of unique values = 1	Remove columns which match strategy	EmployeeCount, Over18, Standard Hours
Columns with only 0s and NAs	Columns where count of unique value = 2 and it is only 0s and NAs	Remove columns which match strategy	Some dates (which are holidays) have in_time, out_time as NA. "2015.01.01" "2015.01.14" "2015.01.26" "2015.03.05" "2015.05.01" "2015.07.17" "2015.09.17" "2015.10.02" "2015.11.09" "2015.11.10" "2015.11.11" "2015.12.25"
NA treatment	TotalWorkingYears – Impute value based on currentAge minus the mean age when people start working.	Mean age of people when they start working = 26 years	1 column updated
	NumCompaniesWorked , EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance  No relationship/pattern found with other columns. Removed rows which have NA.	AttritionYes removed = 2.0% AttritionNo Removed = 2.3%  Total rows removed = 102 rows ~ 2.3% of total data.	4 column updated

Final dimensions for data of all employees will all columns merged = 4308 x 28



## Categorical Features

S.No	Feature Name	Remarks
1	Attrition	Possible values = <i>Yes/No</i>
2	BusinessTravel	Possible values = <i>Non-Travel, Travel_Frequently, Travel_Rarely</i>
3	Department	Possible values = <i>Human Resources, Research &amp; Development, Sales</i>
4	EducationField	Possible values = <i>Human Resources, Life Sciences, Marketing, Medical, Other, Technical Degree</i>
5	Gender	Possible values = <i>Female, Male</i>
6	JobRole	Possible values = <i>Healthcare Representative, Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, Sales Representative</i>
7	MaritalStatus	Possible values = <i>Divorced, Married, Single</i>
8	StockOptionLevel	Possible values = <i>0, 1, 2, 3</i>

## Numerical Features

S.No	Feature Name	Remarks
1	totalOffs, percYearlyWorking, Age, DistanceFromHome	Numeric/continuous values
2	Education	Ordered Categorical – higher number means better education (1 – belowCollege, 5 – Doctor/PhD)
3	JobLevel	Ordered Categorical – higher number means better higher employee band
4	MonthlyIncome, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromot, YearsWithCurrManager	Numerical/continuous values
5	EnvironmentalSatisfaction, JobSatisfaction, WorkLifeBalance, JobInvolvement, PercformanceRating	Ordered Categorical



Frequent Business Travel has a higher percentage of Attrition (~ 25%).

The HR department in itself has a high Attrition rate (~29%). HR is a small department (187 employees) out of which 55 left during the last year.

As reflected in the Department category, Human Resource education field is prone to highest attrition (~40%)

Attrition among Males/Females seem to be balanced.  
Where as "Single" employees, have a high Attrition rate (~25%)

Across different Job Roles, the attrition hovers around company average level, except Research Director Role (~23%).



Attrition rate is same across different stockOptions levels.

**Plot Methodology** : Segmented Column plots  
**Aggregation** : Various (as per plot)



# Univariate & Segmented Univariate – Numerical Features Analysis (2/4)

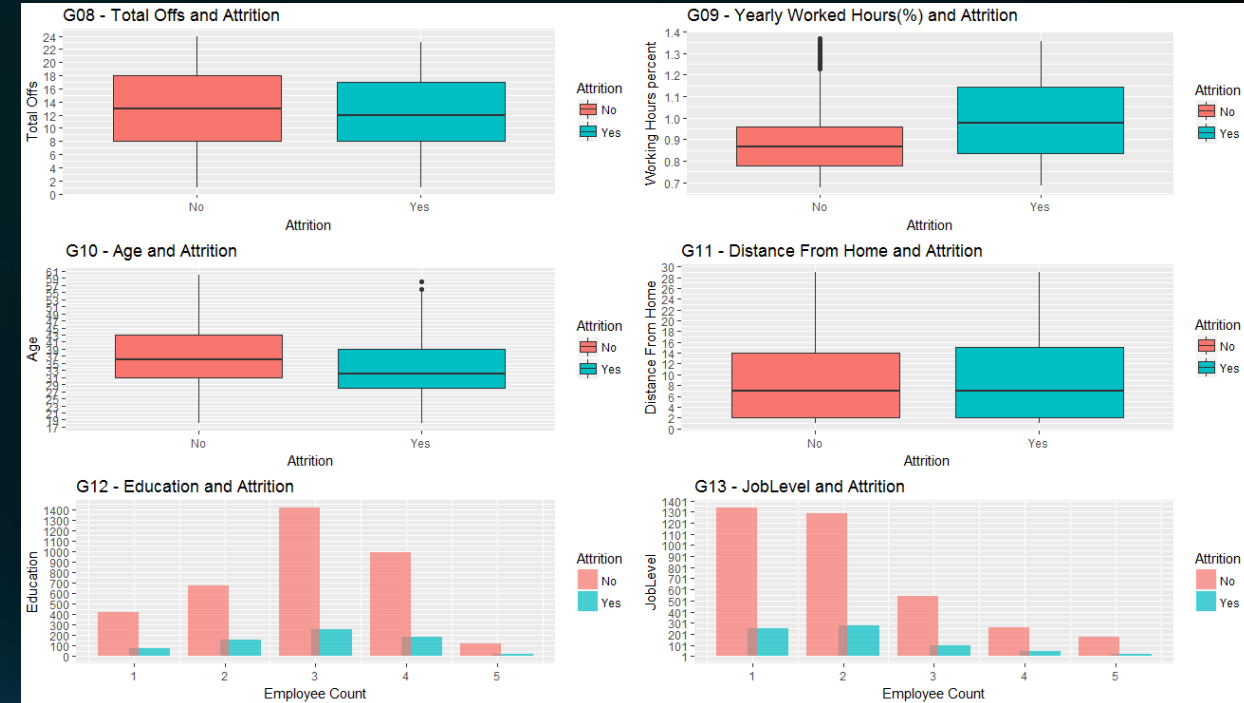
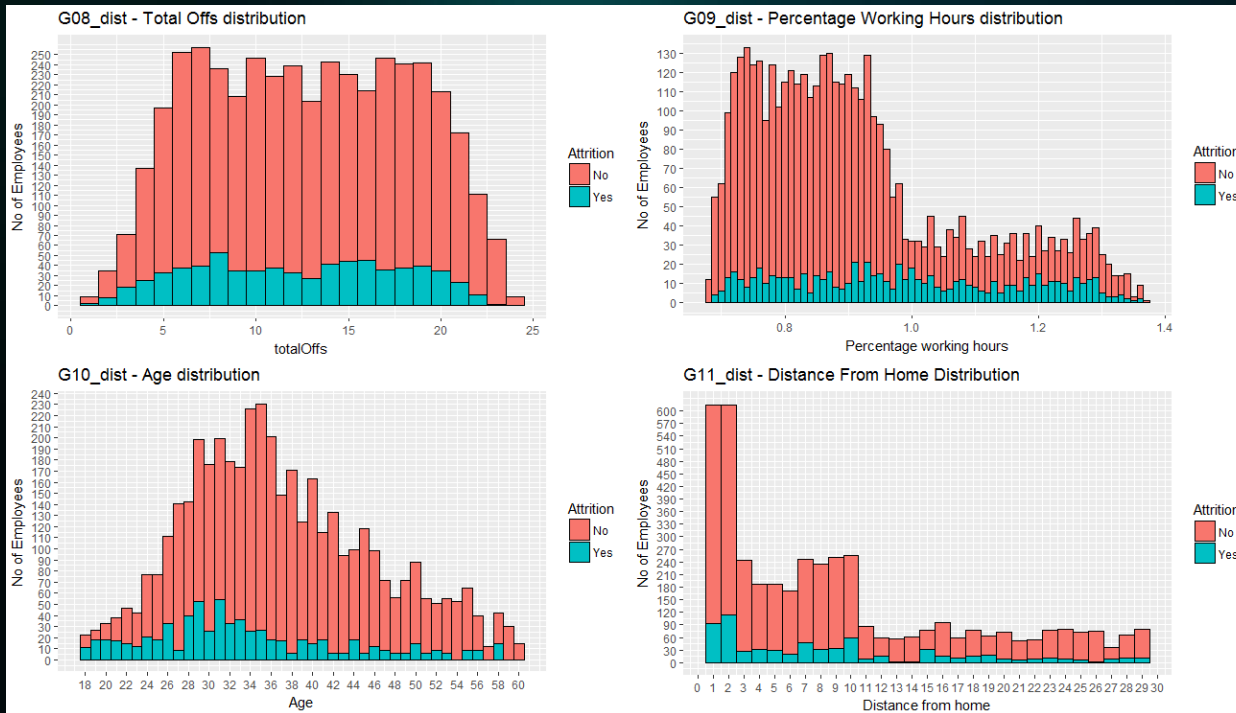
## Inference :

G(08) – TotalOffs & G(11) – DistanceFromHome

Attrition based on TotalOffs has similar distribution across different number of offs. DistanceFromHome presents a mixed pattern about attrition.

G(10) – PercentageWorkingHours & G(10) - Age

Higher Percentage working hours clearly have higher attrition rate. Attrition among employees with age < 32 years is higher as compared to senior employees



**Plot Methodology**  
Aggregation

: Segmented Column plots & boxplots  
: Various (as per plot)

## Inference :

G(12)

Education (1 = BelowCollege, 5 = Doctor/PhD) has similar attrition rate across different levels (~ 15-18%)

G(13)

JobLevel, also has similar attrition rate across different levels.

# Univariate & Segmented Univariate – Numerical Features Analysis (3/4)



## Inference :

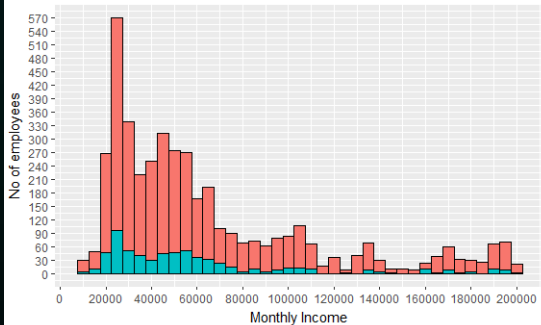
### G(14) – MonthlyIncome

Lower monthly income has higher attrition rate, and the attrition rate reduces as the income increases.

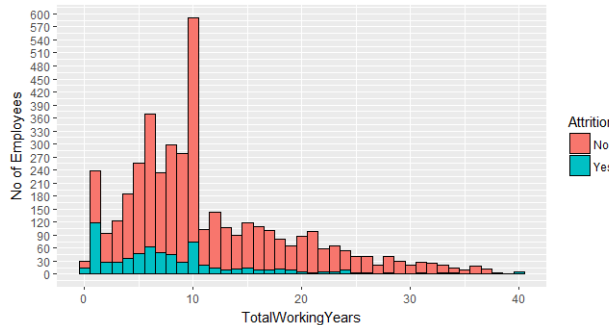
### G(20) – Total Experience & G(10) – Experience at XYZ Corp

Very high attrition rate for employees with just 1 year of experience (Freshers). Interestingly, people who have spend 2 years at XYZ Corp have a higher attrition rate.

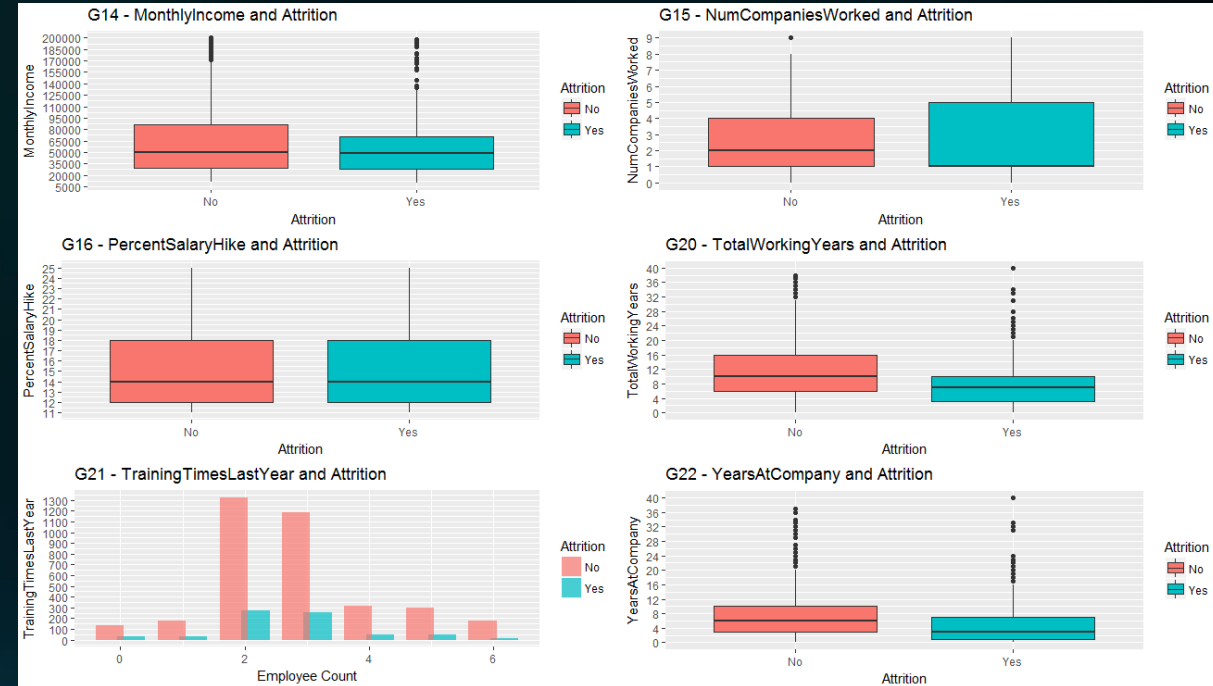
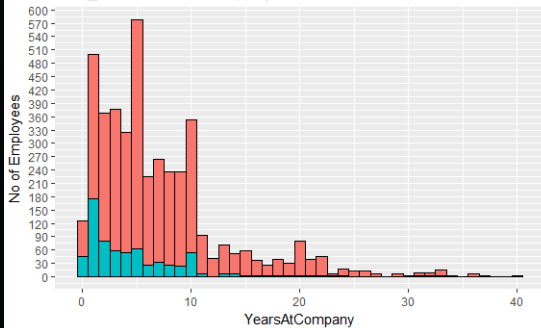
G14\_hist - Monthly Income Distribution



G20\_hist - TotalWorkingYears



G22\_hist - YearsAtCompany



**Plot Methodology** : Segmented Column plots  
**Aggregation** : Various (as per plot)

## Inference :

### G(21) – Trainings in last year

Employees who has 0,2 & 3 training sessions in the last year have a higher attrition rate, rather than the ones who had 1 or more than 3 training sessions.

It seems no training or only some training is not appreciated by employees. This information may also be considered as a reflection of effectiveness of training.

# Univariate & Segmented Univariate – Numerical Features (Contd)

## Analysis (4/4)



### Inference :

#### G(25) – Environment Satisfaction

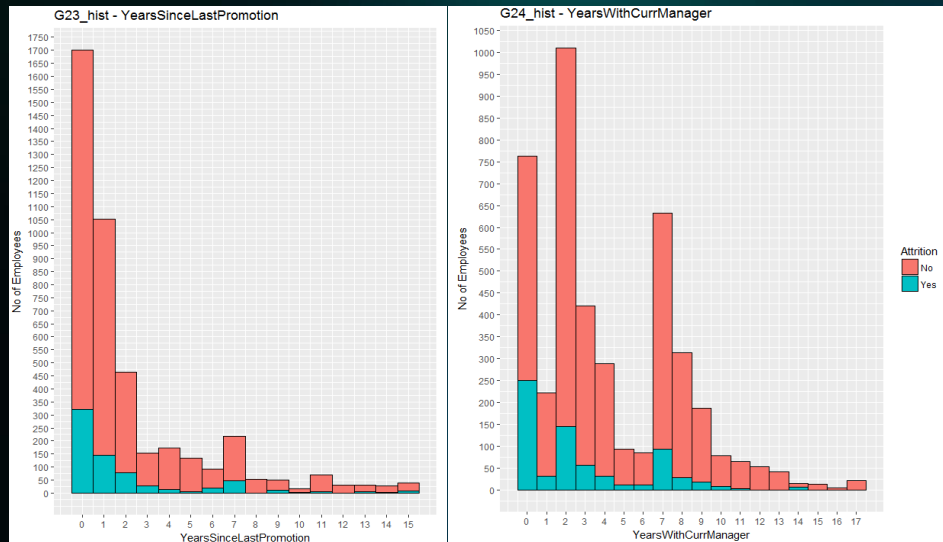
Attrition rate is very high (~ 25%) when Environment Satisfaction = 1. Attrition percentage reduces as environment satisfaction increases.

#### G(26) – Job Satisfaction

Job Satisfaction seem to be highly influential parameter. There is clear decrease in attrition as job satisfaction increases, which is an obvious reason.

#### G(23) – Yrs since Promotion G(24) – Yrs with Same Manager

7 years seems to be an important crucial year both for Promotion and same Manager perspective. Attrition rate seems to be high at 7<sup>th</sup> year for both categories.



### Inference :

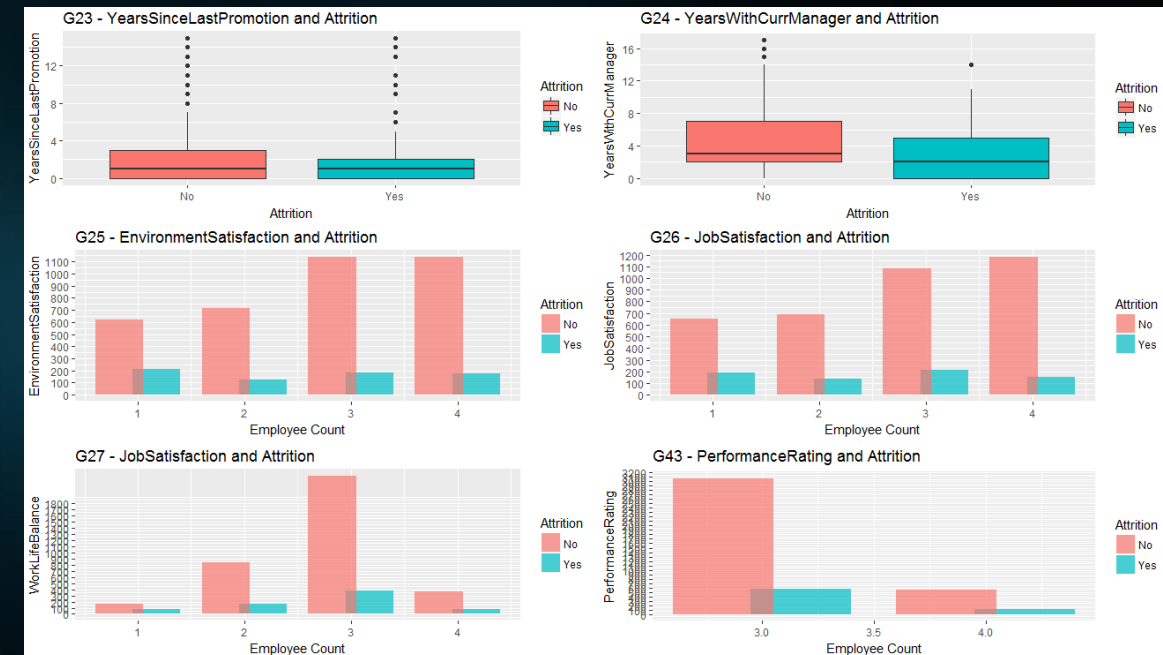
#### G(27) – Work Life Balance

Work life balance by obvious reasons, has high influence on attrition. Work life balance =1 equates to around 30% attrition.

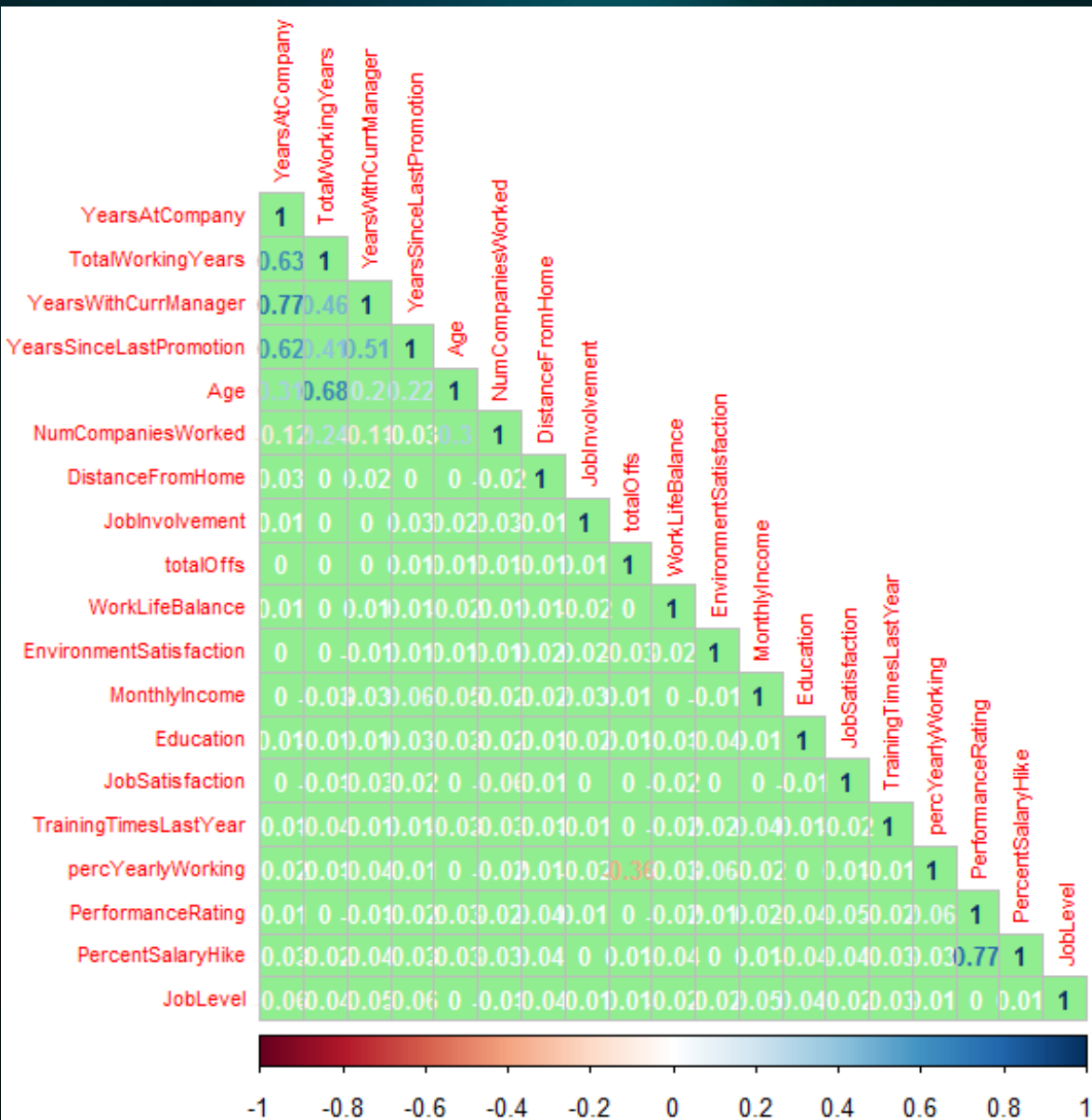
#### G(27) – Performance Rating

Interestingly employees do not heave much importance to performance rating that managers give and hence attrition seems not effected.

**Plot Methodology** : Segmented Column plots  
**Aggregation** : Various (as per plot)



# Bivariate Analysis - Correlation between numerical variables



## Fundamental of correlation in regression:

- One of the objectives is to identify significant variables which effect the business problem. Patterns in highly correlated variables will be similar and will not lead to better insights.
- Fundamentally, regression based models work on the principle of

$$\text{Dependent} = \text{coefficient} * \text{independentVar} + \text{constant}$$

Having co-related features defeat the purpose of independent variables

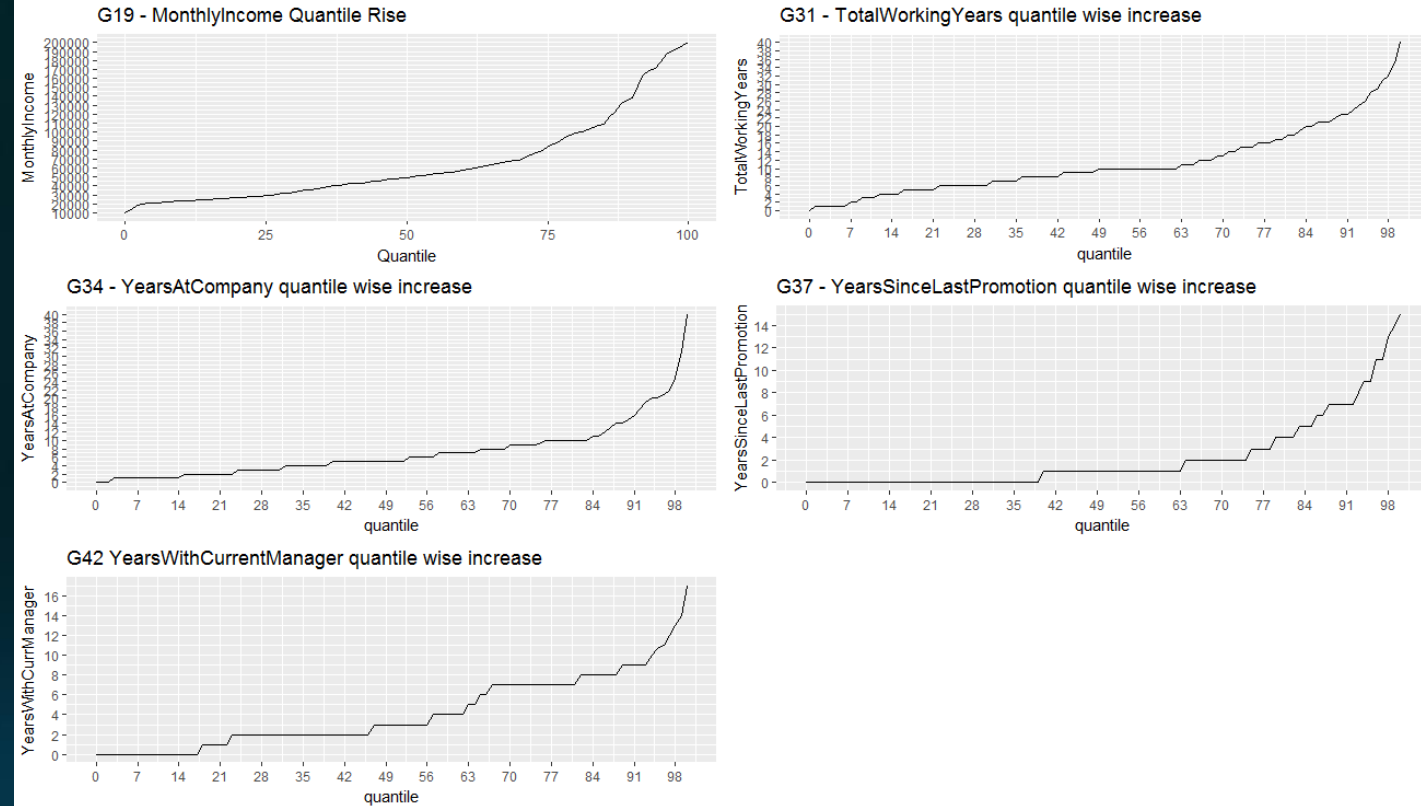
- Having multiple correlated variables in the model, will undermine the detection and prediction capability of the model as it would require more number of iterations to reach the global minima during iterative regression approach.

## HR Case Study Data set (merged)

There is no significant/strong correlation seen between any 2 numerical variables. However, there are chances that correlation may be found when we evaluate 3 or more variables together. Such an analysis will be taken up during stepAIC evaluation.

## Outlier Treatment rule employed

- Various numerical features taken into consideration
  - MonthlyIncome
  - TotalWorkingYears
  - YearsAtCompany
  - YearsSinceLastPromotion
  - YearsWithCurrentManager
- For every feature, a quantile & boxplot is taken for reference.
- UpperEdge chosen for performing capping of outliers should result in affecting less 7-10% of complete dataset.



MonthlyIncome Outlier Treatment		
Upper Edge	No. of rows	Perc of rows
Rs.1,70,000	282	6.5%
YearsAtCompany Outlier Treatment		
Upper Edge	No. of rows	Perc of rows
22 years	111	2.57%

TotalWorkingYears Outlier Treatment		
Upper Edge	No. of rows	Perc of rows
26 years	247	5.7%
YearsSincePromotion Outlier Treatment		
Upper Edge	No. of rows	Perc of rows
11 years	126	2.92%

YearsWithCurrManger Outlier Treatment		
Upper Edge	No. of rows	Perc of rows
11 years	151	3.5%



## Information related to data used for model building

Number of numerical features	19
Number of categorical columns used	7 (Including Gender)
Columns for model.matrix functionality	6 (excluding Gender)
No. of columns made after model.matrix	22
Final Independent column (X)	Numerical + model.matrix + Gender + 1 information column (employee ID)  19 + 22 + 1 + 1 = 42 columns
Dependent Column	"Attrition"
Final Dimensions of data frame used for modeling = 4308 X 44 (train = 70% of 4308 & test = 30% of 4308)	

Operation	AIC value
Model 1 (42 independent + 1 dependent)	2118.4
StepAIC execution Number of columns suggested : 25	2092.36
Model 13 (Final Model) Number of significant columns : 13 (excluding intercept)	2117.4

## Features for model building

totalOffs	percYearlyWorking	Age
DistanceFromHome	Education	JobLevel
MonthlyIncome	NumCompaniesWorked	PercentSalaryHike
TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany
YearsSinceLastPromotion	YearsWithCurrManager	EnvironmentSatisfaction
JobSatisfaction	WorkLifeBalance	JobInvolvement
PerformanceRating	Gender	BusinessTravel.xTravel_Frequently
BusinessTravel.xTravel_Rarely	Department.xResearch...Development	Department.xSales
EducationField.xLife.Sciences	EducationField.xMarketing	EducationField.xMedical
EducationField.xOther	EducationField.xTechnical.Degree	JobRole.xHuman.Resources
JobRole.xLaboratory.Technician	JobRole.xManager	JobRole.xManufacturing.Director
JobRole.xResearch.Director	JobRole.xResearch.Scientist	JobRole.xSales.Executive
JobRole.xSales.Representative	MaritalStatus.xMarried	MaritalStatus.xSingle
StockOptionLevel.x1	StockOptionLevel.x2	StockOptionLevel.x3



# Significant Features

## P-values and coefficients

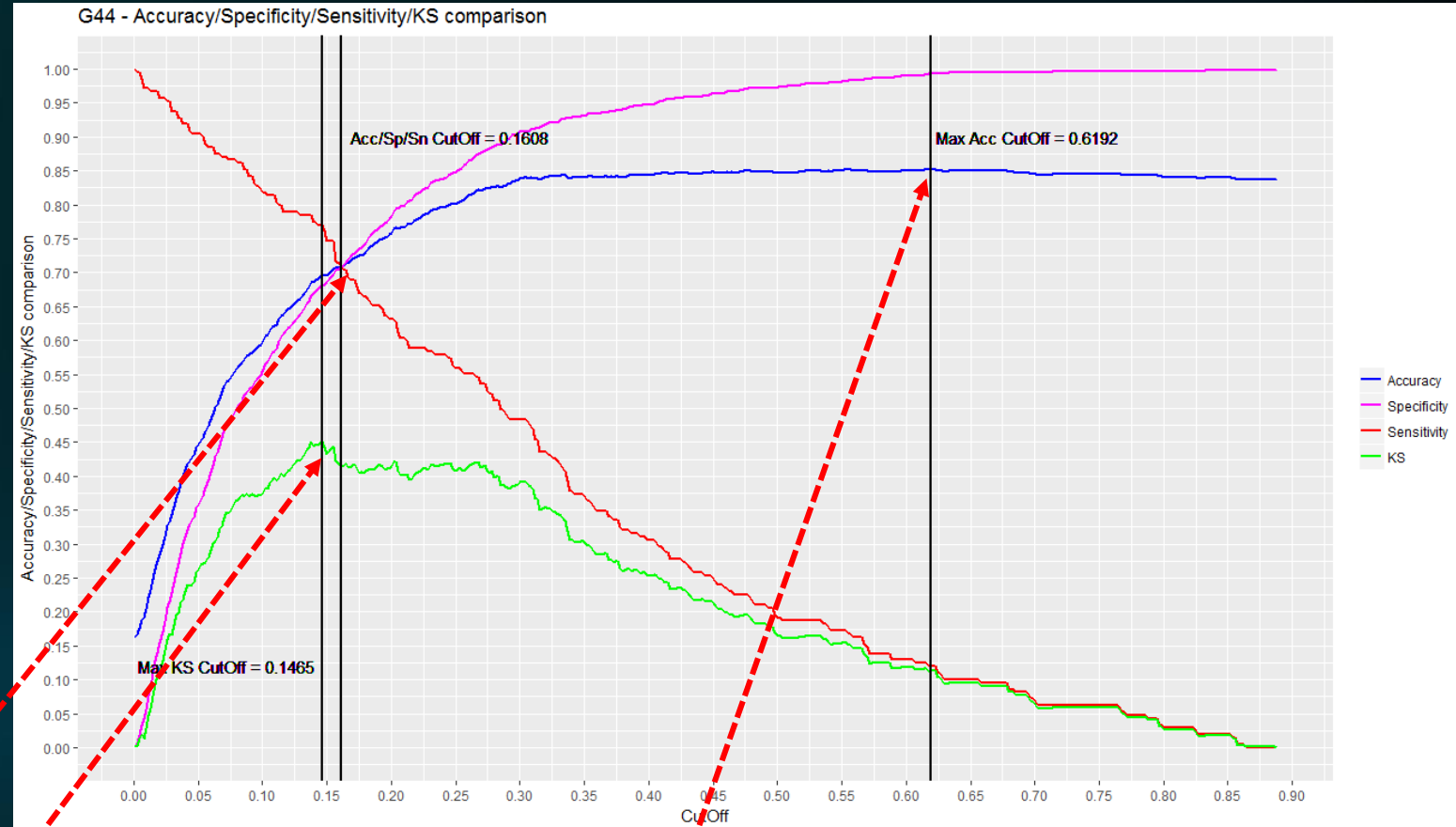
Feature	Coefficient	P-value	Significance Stars
(Intercept)	-2.6219	2.00E-16	***
percYearlyWorking	0.58747	2.00E-16	***
Age	-0.26054	0.00092	***
NumCompaniesWorked	0.36017	3.56E-10	***
TotalWorkingYears	-0.64254	7.79E-10	***
TrainingTimesLastYear	-0.20913	0.00028	***
YearsSinceLastPromotion	0.62684	6.19E-13	***

Feature	Coefficient	P-value	Significance Stars
YearsWithCurrManager	-0.5809	8.29E-10	***
EnvironmentSatisfaction	-0.38163	6.03E-12	***
JobSatisfaction	-0.39377	3.16E-12	***
WorkLifeBalance	-0.24788	5.64E-06	***
BusinessTravel.xTravel_Frequently	0.78395	1.66E-09	***
JobRole.xManufacturing.Director	-0.92506	3.21E-05	***
MaritalStatus.xSingle	1.03141	2.00E-16	***

*Note: A logical understanding of the above co-efficient and their impact on attrition is given at the end of the presentation*

# Model Evaluation – Deciding different cut offs (1 of 4)

- ❑ Step 1- Once Model preparation is completed a cut off value needs to be decided for classifying 1s and 0s for test data set. Following 3 metrics have been used to decide 3 different cut offs:
  - ❑ Cut off at intersection of Accuracy, Sensitivity and Specificity- This will give a balanced model.
  - ❑ Cut off with Highest accuracy- Gives the highest accuracy but model may be unbalanced.
  - ❑ Cut off with highest KS value- KS metric calculates separation between classes. Maximum value of KS is taken to get the greatest separation. Gives a good balance between Accuracy and class separation.






*Cutoff value at Accuracy/SN/SP intersection  
(cutoff = 0.1608)*

*Cutoff with maximum KS (cutoff = 0.1465)*

*Cutoff with maximum Accuracy (cutoff = 0.6192)*

# Model Evaluation – Choosing the best cutoff (2 of 4)

- ❑ Step 2- Once cutoffs have been chosen a best cutoff needs to be decided based on a combination of Statistical evaluation criteria and Business goals:
  - ❖ Statistical Criteria- Ideally model should have high Accuracy, Sensitivity and Specificity along with high AUC (area under the curve).
  - ❖ Business Goals- Our primary goal is to predict correctly the people who are mostly likely to resign (1s in the current context). Even if such a model has lower accuracy it can be acceptable as it will identify employees likely to resign and enable HR of XYZ company to take preventive action.

Criterion	Cutoff Value	Accuracy	Sensitivity	Specificity	Confusion Matrix	Comments
 Intersection of Acc/SN/SP	0.1608	0.707	0.707	0.708	<pre> #               Reference # Prediction    0      1 #              0  766  61 #              1  317 148                     </pre>	1) Balanced Model 2) Misclassifies few 1s as 0s 3) Does not match business need
 Maximum Accuracy	0.6192	0.852	0.119	0.993	<pre> #               Reference # Prediction    0      1 #              0 1076 184 #              1    7   25                     </pre>	1) High Accuracy and Specificity but dismal sensitivity 2) Highly unbalanced
 Maximum KS value	0.1465	0.695	0.770	0.680	<pre> #               Reference # Prediction    0      1 #              0  737  48 #              1  346 161                     </pre>	1) Low Accuracy but balanced Specificity and Sensitivity 2) Best model for Business objective

# Model Evaluation Metrics- ROC curves and AUC (3 of 4)

- ROC curves are line plots of Sensitivity against Specificity

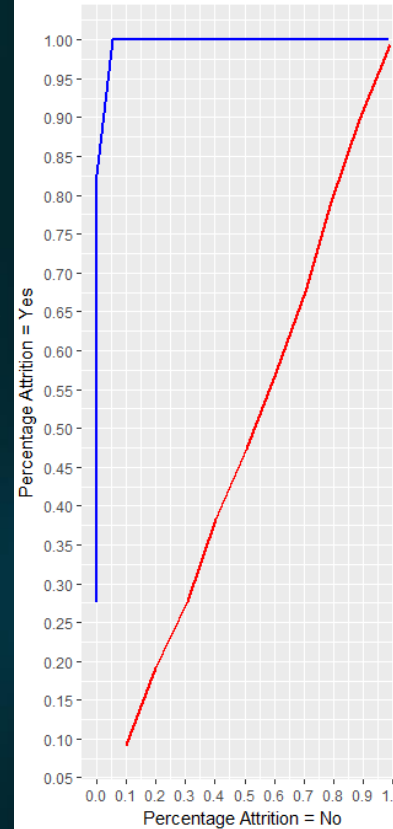
- ❖ An ideal ROC Curve should be nearly a rectangle between the axes
- ❖ While the worst ROC curve will be a 45° line between the axes

ROC Curves for our Model based on different Cut offs

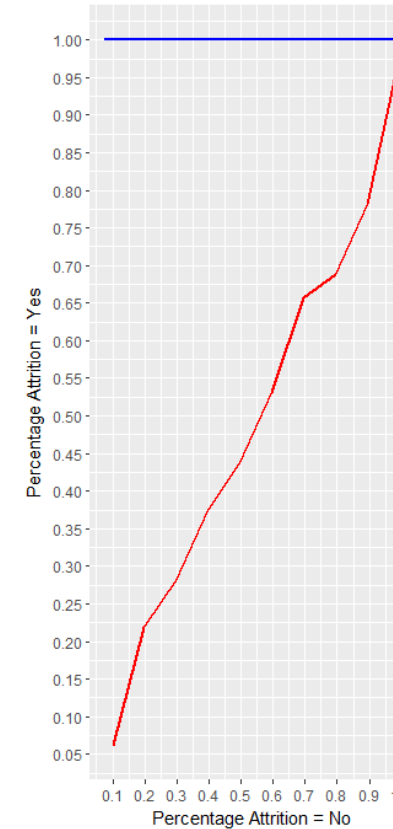
## Inference :

- ❖ Cutoffs 1 and 3 are able to achieve a good separation between classes as can be seen from the almost rectangular ROC curves
- ❖ The ROC curve for KS-Statistics cut off is slightly sharper (higher slope)
- ❖ KS should be our metric of choice as it achieves a slightly sharper curve and has greater AUC. (AUC available in next slide)

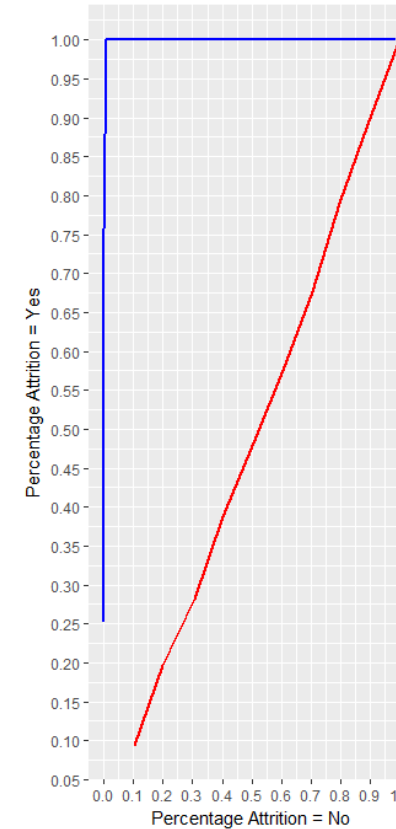
G45 - ROC Curve for Prediction based on Accuracy, Specificity, Sensitivity  
CutOff = 0.1608



G46 - ROC Curve for Prediction based on Max Accuracy  
CutOff = 0.6192

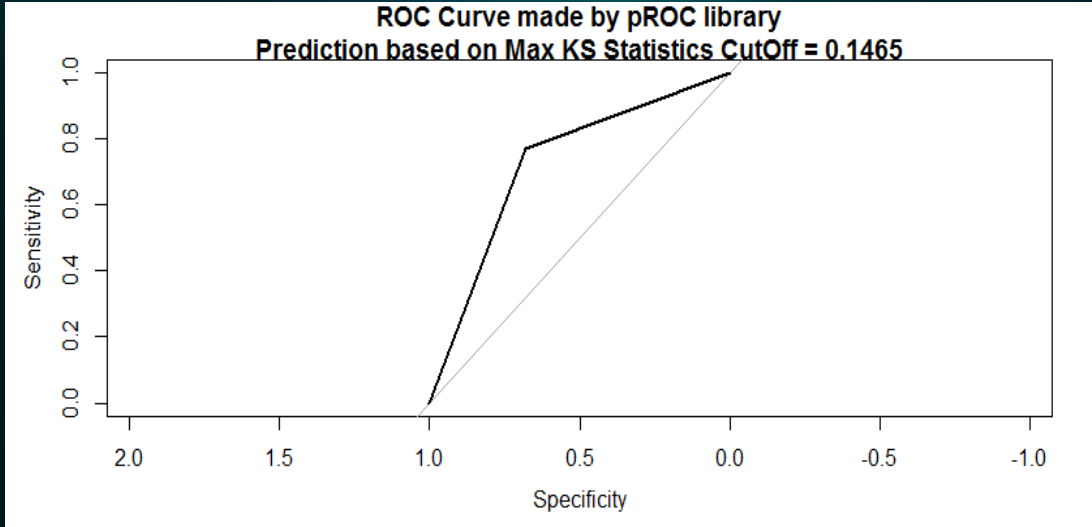


G47 - ROC Curve for Prediction based on KS-Statistics  
CutOff = 0.1465

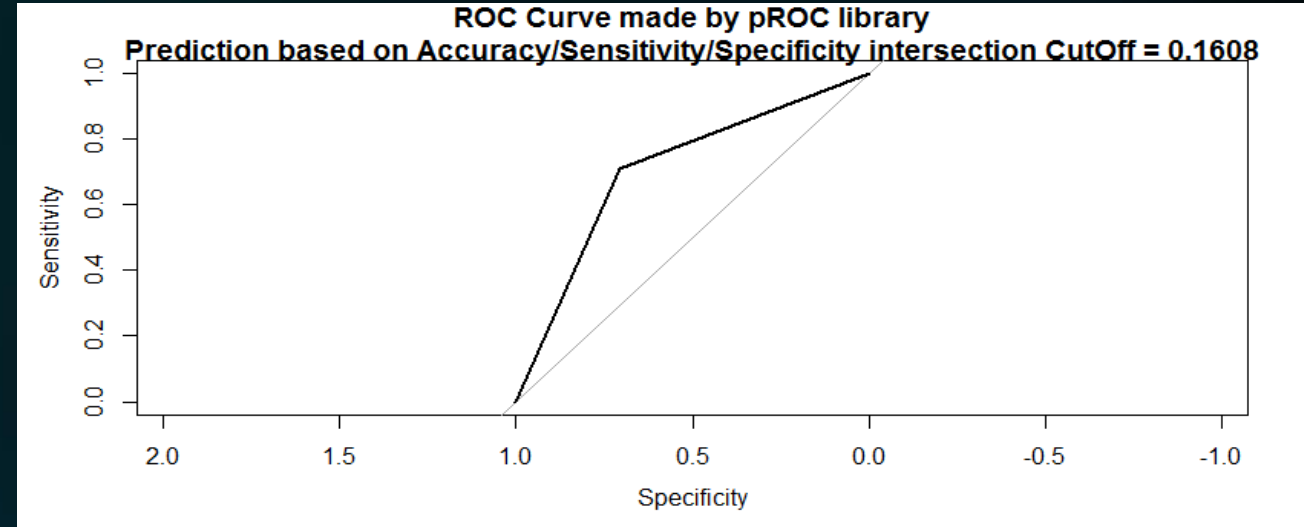


# Model Evaluation Metrics- ROC curves and AUC (4 of 4)

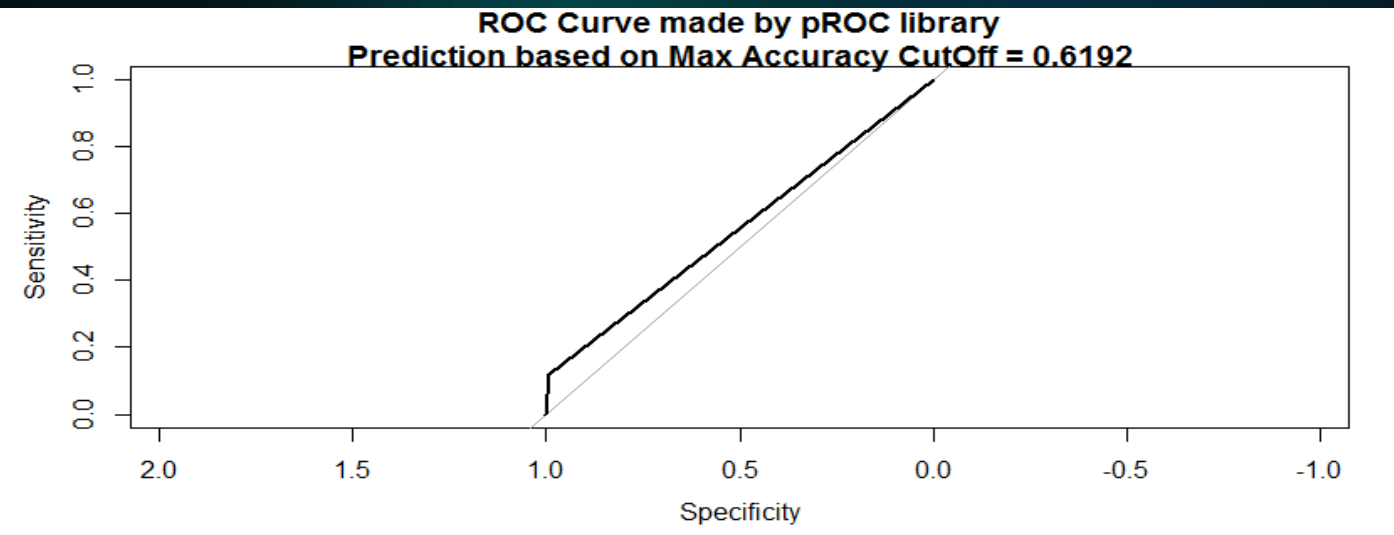
**ROC curve for KS cutoff (AUC=0.7254)**



**ROC curve for Accuracy/Sensitivity/Specificity cutoff (AUC=0.7077)**



**ROC curve for maximum Accuracy cutoff (AUC=0.5566)**



## Inference

- ❑ KS curve gives slightly higher AUC.
- ❑ KS also detects 1s better as no. of ones predicted as 0s is lower. (Ref Slide : 17)
- ❑ Though it has a comparatively low Accuracy it is the most suitable metric for our current business objective as it is able to tap fairly good number of employees who are most probable to resign.

# Inferences From Significant Features

Feature	Coefficient	Significance
percYearlyWorking	0.58747	Positive coefficient : More working hours = higher attrition probability
Age	-0.26054	Negative coefficient : Higher the age = lesser attrition probability
NumCompaniesWorked	0.36017	Positive coefficient : Employee who switches job frequently = higher attrition probability
TotalWorkingYears	-0.64254	Negative coefficient : Higher the total experience of employee = lesser the attrition probability
TrainingTimesLastYear	-0.20913	Negative coefficient : More the trainings = lesser is the attrition probability
YearsSinceLastPromotion	0.62684	Positive coefficient : Later the promotion = higher attrition probability
YearsWithCurrManager	-0.58090	Negative coefficient : Changing the manager frequently = higher attrition probability
EnvironmentSatisfaction	-0.38163	Negative coefficient : Higher the env satisfaction = lesser is attrition probability
JobSatisfaction	-0.39377	Negative coefficient : Higher is the job satisfaction = lesser is the attrition probability
WorkLifeBalance	-0.24788	Negative coefficient : Higher is work life balance = lesser is the attrition probability
BusinessTravel.xTravel_Frequently	0.78395	Positive coefficient : More business trips = higher attrition probability
JobRole.xManufacturing.Director	-0.92506	Negative coefficient : Higher job role = lesser attrition probability
MaritalStatus.xSingle	1.03141	Positive Coefficient : Single employees = higher attrition probability

Note : All significant variables modeled and their coefficients and inferences stated above are logically sound and fall in line with the business issue discussed in this case study. Further analysis and recommendations are stated in next slide



## Important Recommendations

- ❑ One of the primary and highly influential factors are the number of hours the employee is putting in. XYZ Corporation should plan to avoid late working hours for employees. This would require better project planning by management.
- ❑ Promotions has also come out as a important factor influencing attrition. Promotions & appreciation should become a core value for the organization. This will help keep employee motivation high and lower the probability of attrition.
- ❑ Changing of managers very frequently is a negative element for attrition. It seems employees have concern and reservation against working with new managers or different managers very frequently.
- ❑ Job Satisfaction, Environmental Satisfaction & Work Life Balance, are obvious reasons influencing attrition. HR should work to enhance employee engagement apart from regular work by organizing events and team outings. Team building activities would also help increase employee satisfaction.
- ❑ Business Travel has come out to be negative element for attrition. Higher management should leverage technologically enhanced tools (WebEx, Virtualization and Remote Management etc) for avoiding travel for employees.



The above are the prime factors influencing attrition. If XYZ Corporation is able to work on important parameters and recommendation stated above, attrition rate at the company can be controlled.

**Thank You!**